強化学習による２アクチュエータ ５リンク環状ロボットの移動動作獲得

○荒牧 岳志 (東京工業大学)　木村 元 (東京工業大学)
小俣 透 (東京工業大学)　小林 重信 (東京工業大学)

A Reinforcement Learning of the 5-Links Ring Robot with two actuators

*Takeshi ARAMAKI, Hajime KIMURA, Toru OMATA, Shigenobu KOBAYASHI,
Tokyo Institute of Technology.

Abstract － Recently, a control method of a movement of the robot that consists of circular five-links is expected. The robot changes the form by using servomotors attaching two joints, and goes ahead. But it is difficult for designers to build a controller, because the model has closed links and asymmetry and nonlinear. We apply reinforcement learning to the problem. It is a self-adaptive learning and obtains control rules through interaction with the environment. We choose Actor-Critic as its algorithm. Before a control of the real robot, we simulated it. And it can successfully learn to go ahead.

Key Words: Reinforcement Learning, 5-Links Ring Robot, Somersault

1. はじめに

現在、5つのリンクを環状に結んだロボット(Fig.2)の前進動作のより少ないアクチュエータを用いた制御手法が求められている[1)2)]。ロボットは2つの関節に付けられている角度制御のサーボモータにより形を変え、でんぐり返しを行うことにより前に進む事ができる。

ところが、このような閉リンク機構で、非線形非対称のモデルでは、人間が従来の制御手法により設計する事は困難である。そこで、適応的な制御規則の獲得法として知られる「強化学習」[3)]を用いる。強化学習は試行錯誤を通じて環境に適応する学習制御の枠組であり、教師付き学習(supervised learning)とは異なり、状態入力に対する正しい行動出力を明示的に示す教師が存在しない。その代わりに報酬というスカラーの情報を手がかりにして学習する。ある状態よりエージェント(コントローラ)がある行動を決め、それに対し報酬が得られる。強化学習は、設計者が複雑なモデルの設計やパラメータの計測が必要ではなく、環境が変化しても適応的な制御を行うので、様々な実問題に対してその有用性が示されている。

強化学習のアルゴリズムには代表的なものにQ-Learning[4)]等があるが、今回は連続値の行動を扱う事ができ、状態観測の不完全性にも強いと言われる確率的傾斜法に基づくActor-Critic[5)6)]を用いる。

実機(Fig.1)での制御の前段階として、シミュレーションでその制御を行う。シミュレーションでは、状態を観測するセンサとして、どのようなものが適切なのかを測るために4つのセンサを比較する。

2. 実験環境

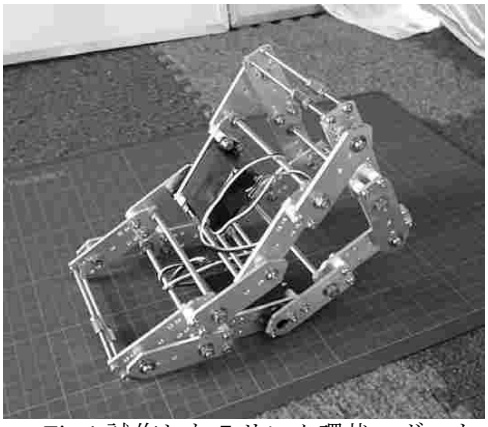

Fig.1 試作した５リンク環状ロボット

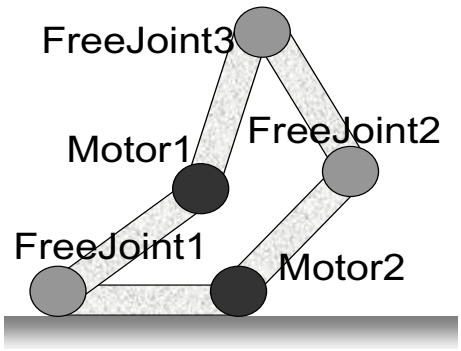

Fig.2　５リンク環状ロボットのモデル図

今回、実装する前に状態を観測するためにどのようなセンサを使用するべきかを調べるために4つのセンサについてシミュレーションする。ただし、以下のセンサはいずれも、On , Offだけを調べる離散的な測定を行う。

1. **傾斜センサ**　水銀スイッチ等で知られる安価なセンサである。5つの各リンクの中央に取りつけ、

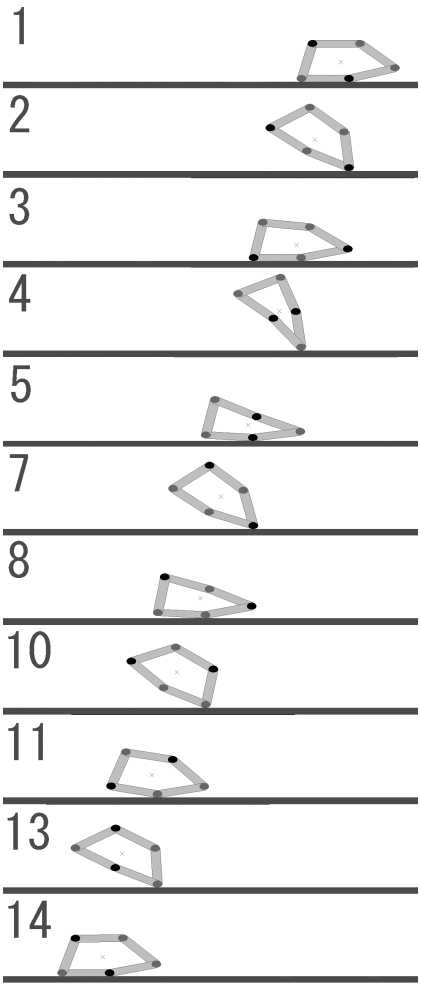

Fig 3.強化学習に得られた左に進む前進動作

そのリンクの角度を調べる。

2. **タッチセンサ**　各関節部に取りつけ、その関節が地面に接地しているかどうか調べる。
3. **光近接センサ**　各リンク中央に取りつけ、そのリンクが地面に接地しているかどうかを調べる。
4. **光近接センサ＋保持回路**　3.のセンサの場合、5つのセンサがいずれもoffとなっている可能性が高いので、全てがoffとなった場合は前にいずれかがonになったときのセンサの状態を示すような状態保持の回路を光近接センサに加える。

強化学習の行動としては、2 つのサーボモータの角度とし､状態は 5 つのセンサを 32 状態に離散化したものと 2 つのサーボモータの角度とする。1step を 300[ms]とし、シミュレーションを行った。

3. 実験結果

Fig.3 は、光近接センサ＋保持回路を用いた時に、強化学習で得られた動きである。十分に学習を終えた時には、1～14 の動きを繰り返し、前(左)に進んでいく。また、4 つのセンサは全て Fig.3 のような動き

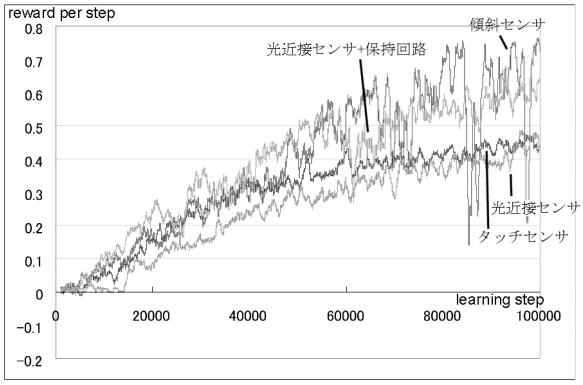

Fig4.各センサを用いた時の学習曲線

を行い、その行動に差は生じなかった。

Fig.4 は、学習曲線を示し、縦軸は 100step 毎の平均報酬値である。どのセンサも一様に学習ができているが、タッチセンサや光近接センサでは、5 つのセンサが全て off になることがあり、うまく状態観測ができず､学習が他の 2 種類に比べて遅くなっている。

4. 考察

- 強化学習で、どのセンサに対しても、制御規則を得ることができた。
- 状態観測の不完全性の少ないセンサ（1.傾斜スイッチと 4.光近接センサ＋保持回路）の方が他の 2 種類に対して、性能がよかった。
- 実時間で学習を行うには､まだ学習に時間がかかりすぎており、新たな工夫を要する。
- 今後は実機で強化学習を実装し、確かめる。

参考文献

1) 森 治、小峰 晋一郎、村木 誠一郎、小俣 透：劣駆動結合によるパラレルメカニズムへの形態変化：垂直面内での劣駆動シリアルリンク系の終端制御、本学術講演会
2) Vunthichai Ampornaramveth：On Motion Generation of Autonomous Decentralized Mechanical Systems Using Genetic Methods, ph. D. thesis, Tokyo Institute of Technology (1999)
3) R.S. Sutton and A.G. Barto：Reinforcement Learning , An Introduction. A Bradford Book. The MIT Press (1998)
4) C.J.C.H Watkins and P.Dayan：Technical note : Q-Learning, Machine Learning, Vol. 8, pp. 279-292 (1992)
5) 木村 元, 小林 重信：Actor に適正度の履歴を用いた Actor-Critic アルゴリズム, 不完全な Value Function のもとでの強化学習, 人工知能学会, Vol 15, No.2, pp.267-275(2000)
6) 山下 透, 木村 元, 小林 重信：強化学習による多足歩行ロボットの実現, 計測自動制御学会, 第 13 回自立分散システムシンポジウム資料 pp 111-116(2001)